Proceedings of ICLLCD 2026 Symposium: Using Visual Arts to Enrich History Understanding
DOI: 10.54254/2753-7064/2026.HT31555

Continuous Valence–Arousal Regression for Music Emotion Estimation Using Machine Learning with OpenL3 Embeddings

Jingyi Lyu

Department of Humanities, Communication University of China, Beijing, China
ljy@mails.cuc.edu.cn

Abstract. Music Emotion Recognition (MER) aims to model the mapping between acoustic features and emotional representations. MER has important value in applications such as music recommendation, automatic accompaniment, and automatic music generation, yet remains challenging due to strong subjectivity and complex temporal dynamics. Based on the public MediaEval Database for Emotional Analysis of Music (DEAM), this study employs OpenL3 pre-trained audio embeddings as a unified feature representation and performs frame-level feature extraction with temporal alignment to 2 Hz continuous Valence-Arousal (VA) annotations. On this basis, three lightweight regression models, including a Multilayer Perception (MLP), a Bidirectional Long Short-term Memory network (BiLSTM), and a Transformer Encoder are constructed to VA regression. Model performance is evaluated using Root Square Error (RMSE) and the Pearson Correlation Coefficient (PCC). Experimental results show that the MLP achieves the best overall performance on the test set, with lower RMSE and higher correlation than the BiLSTM and Transformer Encoder. These results demonstrate that pre-trained audio representations enable stable and efficient continuous music emotion regression with lightweight machine learning models.

Keywords: Music emotion recognition, machine learning, valence, arousal, OpenL3

1. Introduction

Music is a fundamental carrier of emotion. Music Emotion Recognition (MER) aims to automatically model the mapping between musical features and affective representations, and has become a central research topic in the field of Music Information Retrieval (MIR) [1]. In recent years, driven by its increasing relevance to applications such as music recommendation, automatic accompaniment, and automatic music generation (AIGC), MER has attracted substantial attention from both academia and industry [2].

The continuous Valence–Arousal (VA) representation has been widely adopted in MER regression tasks due to its ability to characterize the dynamic evolution of musical emotion in a time series manner [3,4]. The publicly available MediaEval Database for Emotional Analysis of Music (DEAM) provides 1,802 music excerpts with corresponding continuous VA annotations sampled at 2

DOI: 10.54254/2753-7064/2026.HT31555

Hz, and has become a commonly used benchmark for continuous music emotion prediction [5]. As a pre-trained audio representation model, OpenL3 is capable of extracting deep features that capture timbral, rhythmic, and high-level semantic information from audio and audiovisual signals, and has been extensively applied in MER-related analyses [6].

Building upon the continuous VA annotations provided by DEAM and the standardized audio features extracted by OpenL3, this study focuses on continuous music Valence–Arousal prediction and constructs three lightweight regression models: a Multilayer Perceptron (MLP), a Bidirectional Long Short-Term Memory network (BiLSTM), and a Transformer Encoder. These models respectively target frame-level mapping, local temporal dependency modeling, and long-range temporal dependency modeling. Under a unified framework including feature extraction, regression modeling, and performance evaluation, their effectiveness in modeling emotional dynamics in music is systematically compared. This design addresses the lack of systematic evaluation of multiple lightweight models under a unified representation setting in existing studies. Experimental results demonstrate that the MLP achieves the best overall performance on the target task.

The contributions of this work are as follows:

(1) A lightweight and efficient framework for continuous music emotion regression based on the DEAM dataset and OpenL3 representations are established

(2) Under unified data and feature conditions, the performance of MLP, BiLSTM, and Transformer Encoder models is comparatively evaluated

(3) A reusable continuous emotion prediction interface is developed, which can be directly applied to arbitrary audio inputs (.mp3/.wav) to generate continuous VA curves

2. Literature review

2.1. Music emotion modeling

Music emotion modeling is based on two widely acknowledged paradigms: discrete categorical and continuous dimensional. Discrete emotion classification is intuitive but cannot capture both the magnitude and subtle variations, while continuous dimensional models use the Valence–Arousal axes for representing mood states. They therefore enable modelling emotional polarity as well as intensity over time. While contrasted with discrete classification, the VA model is better suited to capturing the continuous emotional reactions produced by music. This makes it the dominant mode of representation in MER research [1].

2.2. Deep representations for music emotion

Methods of MER have traditionally been built on handcrafted acoustic features including rhythm, timbre, and pitch. However, such low and mid-level features cannot efficiently encapsulate intricate musical semantics such as harmony, the context of the audio structure as well as the emotional atmosphere, thus limiting their performance in generalization [7]. With the development of deep learning, pre-trained audio embeddings like OpenL3 have been used to automatically extract higher-level representations. From this base, models such as MLP, Recurrent Neural Networks (RNN/LSTM), and Transformer-based models have been proposed to enable the temporal modelling of music emotion [2,8].

DOI: 10.54254/2753-7064/2026.HT31555

2.3. Dataset and evaluation metrics for continuous VA regression

In continuous VA regression research, DEAM provides 1,802 music excerpts with continuous VA annotations sampled at 2 Hz and has been widely adopted as a benchmark for training and evaluating models' ability to predict dynamic variations in musical emotion [9]. In terms of evaluation metrics, studies commonly adopt Root Mean Square Error (RMSE) and the Pearson Correlation Coefficient (PCC) to measure the magnitude of deviation between predicted values and ground-truth annotations, and the consistency of temporal trends, respectively [10].

3. Method

3.1. Problem definition

In this study, continuous music emotion is constructed as a frame-level time-series regression task. For an input audio signal its, in the deep embedding space, frame-level representations are denoted as

$$X=\{x_t\}_{t=1}^{T},\ x_t\in R^{512} \tag{1}$$

The continuous VA annotation sequence is then shown as

$$Y=\{(v_t,a_t)\}_{t=1}^{T},(v_t,a_t)\in R^{2} \tag{2}$$

serves as the supervisory signal. To learn a mapping function

$$f:R^{T\times 512}\rightarrow R^{T\times 2} \tag{3}$$

Such that the predicted sequence approximates the ground-truth annotations Y at the frame level and, thus, modeling for the continuous temporal evolution of musical emotion [10,11].

3.2. OpenL3 feature extraction

In this study, OpenL3 is used to encode raw music signals into deep embeddings, and delivers a uniform audio feature representation. OpenL3 extracts high-level semantic representations from short audio segments through cross-modal self-supervised contrastive learning, and outputs high-dimensional embedding vectors via convolutional networks to characterize timbral attributes, rhythmic patterns and energy distributions [6]. This approach eschews the reliance on traditional handcrafted features while creating a shared feature space for fair comparison between different regression models.

3.3. Model architecture

To ensure a fair comparison among models, all architectures are built for fair comparison among models and trained and run at the same settings with the same input features, evaluation metrics, and data splits. The model architecture and hyperparameters were chosen from initial validation results, focused on stable convergence and general performance.

The MLP is implemented with two fully connected hidden layers containing 256 and 128 units, respectively. ReLU activation functions are employed, together with a dropout rate of 0.2, and the

DOI: 10.54254/2753-7064/2026.HT31555

model performs regression at the frame level. A two-dimensional linear output layer is used to predict Valence and Arousal values. For temporal modeling, a BiLSTM architecture with two bidirectional layers and a hidden dimension of 128 is adopted, allowing short-term temporal dependencies in the emotion sequence to be effectively captured. A Transformer-based encoder is employed to account for longer-range temporal relationships. The encoder includes two blocks with a model dimension of 256, eight attention heads, and a 512-dimensional feed-forward network. To support sequence modeling, a dropout rate of 0.1 and sinusoidal positional encoding are incorporated.

3.4. Hyperparameter configuration

The training hyperparameters have been set up based on initial validation experiments, to stabilize convergence among different model architectures.

The MLP model is optimized by AdamW with a learning rate of 5×10-4, while the BiLSTM and Transformer Encoder are optimized by Adam optimizer with the same learning rate. The batch sizes are 156 for MLP, 64 for BiLSTM and 32 for Transformer Encoder, respectively. The validation loss is used to select the model, and the best checkpoint is kept for final evaluation on the test set.

4. Experiments and results

4.1. Experiments setup

All data are then split into training, validation and test sets in an 8:1:1 ratio, feature normalization is applied using statistics computed from the training set. As input features, all three models utilize 512-dimensional frame-level embeddings extracted from DEAM audio using OpenL3 and utilize DEAM continuous VA annotations as supervised labels and task labels according to a unified task definition. The models acquire a mapping from the deep audio representations to the continuous value of emotion that leads to a comparable evaluation for different model architectures for the same regression task. Specifically, audio signals sampled at 44.1 kHz are used as input. OpenL3 is set to 512-dimensional embeddings using the music-domain model, a window length of 0.1 s, and a hop size of 0.5 s, ensuring temporal alignment with the 2 Hz VA annotations in DEAM.

4.2. Evaluation metrics

The model performance is evaluated through Root Mean Square Error (RMSE) and Pearson Correlation Coefficient (PCC) which measure the prediction accuracy in the magnitude of the error and stability of the time trend. RMSE is the average frame-level prediction error for VA values, which shows how accurately we predict them, with lower values indicating smaller deviations from our ground-truth [12]. PCC quantifies temporal consistency of predicted sequences to ground-truth annotations along the VA dimensions in order to demonstrate how well the model captures dynamic variations in emotions, where value closer to 1 is associated with a greater agreement [13].

4.3. Quantitative results

As illustrated in Figure 1 and Figure 2, the MLP converges more quickly and exhibits a comparatively stable PCC trajectory on both the Valence and Arousal dimensions. The validation RMSE decreases in a gradual and steady manner, eventually settling at a lower level than that observed for the BiLSTM and Transformer Encoder. In comparison, the two sequence-based models

DOI: 10.54254/2753-7064/2026.HT31555

show noticeably larger variations in their correlation curves. Taken together, the results suggest that, under the same feature representation, frame-level regression is sufficient to achieve stable optimization and reliable fitting for continuous emotion prediction.

The quantitative results on the test set are summarized in Table 1. The MLP produces the lowest RMSE and the highest correlation coefficients across both emotional dimensions, outperforming the BiLSTM and Transformer Encoder in terms of overall prediction quality. A further observation is that all three models obtain higher correlation scores for Arousal than for Valence, indicating that Arousal is generally easier to model in the current experimental setting. Figure 3 provides an illustrative example of the continuous VA trajectories generated by the MLP for a representative music excerpt, showing how emotional values change over time.

Table 1. Performance comparison of the three models on the test set

Model	Test RMSE ↓	Test Pearson r(V) ↑	Test Pearson r (A) ↑
MLP	0.1203	0.878	0.909
BiLSTM	0.1903	0.650	0.809
Transformer	0.1772	0.720	0.829

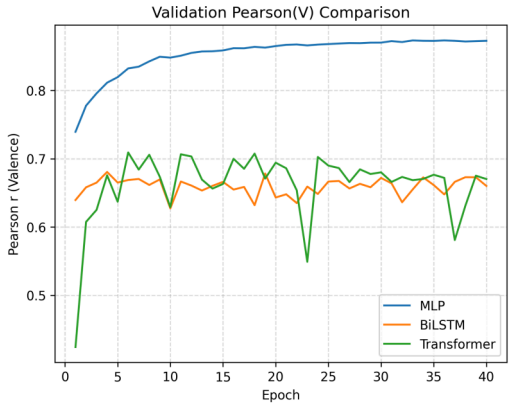

Figure 1. Validation Pearson correlation curves for Valence (picture credit: original)

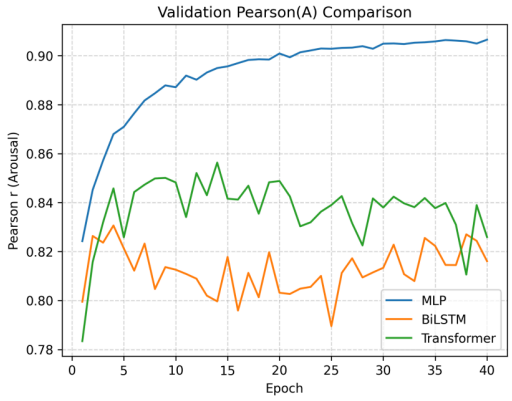

Figure 2. Validation Pearson correlation curves for Arousal (picture credit: original)

DOI: 10.54254/2753-7064/2026.HT31555

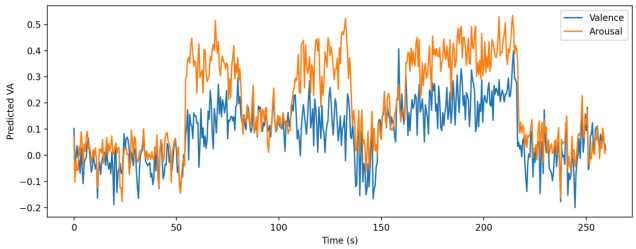

Figure 3. Predicted VA trajectories for a representative song (picture credit: original)

4.4. Discussion

Under a unified task setting, the frame-level MLP achieves the best performance, indicating that OpenL3 features exhibit strong emotion-related relevance for the current task, while the potential advantages of temporal modeling in the BiLSTM and Transformer Encoder are not fully realized. This can be attributed to two factors. On one hand, the continuous emotion annotations in DEAM vary smoothly over time, such that local dynamics can be effectively captured by frame-level features, leaving limited room for complex temporal modeling to provide additional gains. On the other hand, Arousal tends to exhibit stronger correlations with acoustic attributes such as rhythm and energy, which may make its temporal patterns easier for models to learn.

5. Conclusion

This study addresses the task of continuous VA regression in music by constructing three lightweight models – MLP, BiLSTM, and Transformer Encoder – based on the DEAM dataset and OpenL3 pre-trained embeddings, within a unified analysis framework, and then validating them on the test set. Results suggest that, under the same OpenL3 feature representations and the same DEAM data setting, the MLP achieves the best overall performance. This indicates that OpenL3 provides sufficient informative emotion-related representations, enabling simple models to stably reconstruct the dynamic evolution of musical emotion. Such a representation-driven performance paradigm, combining strong feature representations with lightweight models, results in good training efficiency and generalization capability.

In addition, a continuous emotion prediction interface based on the optimal MLP model has been implemented, which has been directly implemented for arbitrary audio input (.mp3/.wav) to compute VA emotion curves, and the present study shows high degree of adaptability and practicality.

Despite these results, the current study is still limited. Future advancements in this area may be developed by integrating multimodal information, i.e., lyrics-melody alignment, to more comprehensively model music-induced emotional experiences. Moreover, exploring the potential of more complex models for long-term emotion dynamics on larger and more diverse datasets is also promising.

References

[1] Han, D., Kong, Y., Han, J. and Wang, G. (2022) A survey of music emotion recognition. Frontiers of Computer Science, 16, 166335.
[2] Jiang, X., Zhang, Y., Lin, G. and Yu, L. (2024) Music emotion recognition based on deep learning: A review. IEEE Access, 12, 157716–157745.
[3] Russell, J.A. (1980) A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1161–1178.

DOI: 10.54254/2753-7064/2026.HT31555

[4] Schaab, L. and Kruspe, A. (2024) Joint sentiment analysis of lyrics and audio in music. arXiv preprint, arXiv: 2405.01988.
[5] Soleymani, M., Aljanaki, A. and Yang, Y.-H. (2018) DEAM: MediaEval database for emotional analysis in music. University of Geneva and Academia Sinica, Switzerland.
[6] Cramer, A., Wu, H.-H., Salamon, J. and Bello, J.P. (2019) Look, listen, and learn more: Design choices for deep audio embeddings. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 3852–3856.
[7] Koh, E. and Dubnov, S. (2021) Comparison and analysis of deep audio embeddings for music emotion recognition. arXiv preprint, arXiv: 2104.06517.
[8] Kang, J. and Herremans, D. (2025) Towards unified music emotion recognition across dimensional and categorical models. arXiv preprint, arXiv: 2502.03979.
[9] Liyanarachchi, R., Joshi, A. and Meijering, E. (2025) A survey on multimodal music emotion recognition. arXiv preprint, arXiv: 2504.18799.
[10] Yang, Y.-H., Lin, Y.-C., Su, Y.-F. and Chen, H.H. (2008) A regression approach to music emotion recognition. IEEE Transactions on Audio, Speech and Language Processing, 16, 448–457.
[11] Chaki, S., Doshi, P., Bhattacharya, S. and Patnaik, P. (2020) Explaining perceived emotion predictions in music: An attentive approach. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 150–156.
[12] Chai, T. and Draxler, R.R. (2014) Root mean square error (RMSE) or mean absolute error (MAE)? — Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7, 1247–1250.
[13] Pearson, K. (1895) Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58, 240–242.